

THE MINIMUM PROFICIENCY IN ENGLISH FOR ENTRY-LEVEL NURSES: TOEFL™
THOMAS R. O'NEILL, PHD | NCSBN TESTING SERVICES

State boards of nursing are charged with protecting the public through the regulation of nursing in their jurisdiction. One of the ways in which they perform that function is by setting and enforcing prerequisite conditions for getting a license. Often these prerequisites are related to education, experience, and demonstrating that they are at least minimally competent by passing the NCLEX®. However, there are some circumstances, such as being educated internationally that might warrant the use of additional prerequisites, such as passing an English proficiency examination.

BACKGROUND

In the years 2001, 2002, and 2003, the number of first-time, NCLEX-RN® candidates who were not educated in one of the NCSBN¹ member board jurisdictions was 8,613, 12,762 and 16,490, respectively. The numbers for first-time, NCLEX-PN® candidates were 1,363, 1,810 and 2,198, respectively. The numbers were even higher when repeat test-takers were included. Clearly, there are a large number of nurse

candidates who were educated outside of the United States and the trend seems to be increasing. For many of these candidates, English is not their primary language. This provides an additional challenge to boards of nursing. Not only do the boards of nursing need information regarding the clinical competence of these candidates, but the board also needs to know if the candidate has adequate language skills to effectively use their clinical skills.

Typically, English language proficiency tests produce a score, not a pass-fail decision. How the score should be interpreted is specific to the purpose. For example, the minimum English proficiency required to be an editor or communications director is likely to be quite different than the minimum proficiency required to be an accountant or actuary. For these reasons, NCSBN set out to establish a recommended minimum standard of English proficiency specific to entry-level nursing. Making available to the member board jurisdictions such a legally defensible passing standard would be an obvious benefit. Rather than have each jurisdiction perform essentially the same study, it seemed sensible to commit significant resources to the project and conduct a single, well-crafted study that could be used by all jurisdictions, should they so choose. This could provide an additional benefit for internationally educated candidates by making the examination results portable across the jurisdictions that use the standard.

¹The National Council of State Boards of Nursing, Inc. (NCSBN) is a not-for-profit organization that is composed of 60 jurisdictional boards of nursing in the United States and US territories whose mission is to provide leadership to advance regulatory excellence for public protection.

Using an underlying principle of public safety, the minimum standard was intended to reflect the level of English language proficiency believed necessary for entry-level nurses to be able to perform important nursing responsibilities safely and effectively. It is recommended that internationally educated nurse-candidates meet or exceed this standard before they are issued a license. It is important to note that the standard was intended to reflect the minimum level of English proficiency necessary for safe and effective entry-level practice, not the level of proficiency necessary for nurse-candidates to take the NCLEX examination.

Before a minimum English proficiency standard can be set, at least one instrument to measure it must be identified. The *Profiles of Member Boards 2002* (NCSBN, 2003) suggests that the most commonly used English proficiency examination used by boards of nursing is the Test of English as a Foreign Language (TOEFL™). Therefore, as a first step in establishing a minimum English proficiency standard, a standard setting study was conducted using the TOEFL examination. The intended examinee population consists of nurse-candidates who have been educated outside of the United States and in a language other than English. English is probably not the first or primary language of these candidates. This population typically includes both inexperienced practitioners and experienced practitioners; regardless, all are seeking entrance into the nursing profession in the United States. For this reason, the minimum level of English language proficiency is understood in the context of entry-level [United States-based] nursing practice.

Standard Setting Process

Although psychometrics can provide useful information for standard setting, the setting of standards is not really a measurement issue. To illustrate this point, consider the ruler. The ruler has been around

for a long time and is generally regarded as a stable instrument for measuring distance. However when a child goes to an amusement park and asks why one must be a certain height to ride a particular ride, the explanation about the ruler's stability seems quite irrelevant. Why not an inch lower? Or higher? Of course, there is a safety-based rationale that considers acceptable risks behind the rule, but how safe a ride should be and what constitutes an acceptable risk are really personal judgments made by a person or a group of people. To implement this judgment evenly across all people, it is necessary to develop a policy. Although psychometrics as a field has been quite successful in devising tests and questionnaires to measure traits, aptitudes, and attitudes that are demonstrably reliable and valid, the selection of cut-off criteria for making classification decisions is not quite as scientific. Its creation is more of a policy decision than a measurement decision. Cizek (2001) expressed this perspective in his book on standard setting.

Although psychometrics falls more along the lines of science, standard setting falls more into the social. Standard setting is the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other. (p. 5)

This perspective, however, has not been universally embraced. Some researchers view the ideal standard as the threshold that optimally classifies candidates with regard to a very specific predicted outcome. From this perspective, one can evaluate the number of correct classifications and the number of false positives and false negatives. Nevertheless, what constitutes an acceptable percentage of false positives and false negatives remains a judgment. A passing standard seems conceptually to be a function of informed professional judgment. There is no passing standard that is empirically correct. A passing score reflects the values of those profes-

sionals who participate in its definition and adoption, and different professionals may hold different sets of values. Its determination may be informed by empirical information or data, but ultimately, the passing standard is a judgment-based decision.

Regardless of one's theoretical perspective, the standard used to classify examinees must not be made in an arbitrary and capricious manner. Furthermore, the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), recommend that the rationale and procedures used to set the standard be clearly documented. This includes a description of the standard setting procedure, the panelist selection process and the qualifications of panelists selected, as well as, a description of the training provided. This report documents these aspects of the standard setting process.

METHODS

TOEFL examination

The TOEFL is an examination designed to assess English language ability in examinees for whom English is not their native language. With the exception of the writing section, all items are dichotomously scored and use a Selected Response (SR) type item format (essentially multiple-choice-question format, although the mechanism for selection can vary a bit and multiple responses are required in some instances). The writing sample is a single prompt that is scored 0-6. Three section-level subscores (0 to 30 points each) and a total TOEFL score (0 to 300 points) are reported. The total TOEFL score is the average of the three section scores multiplied by 10. The three sections are: (1) Listening Comprehension, (2) Structure and Written Expression, and (3) Reading Comprehension.

Standard Setting Methods

Two standard setting procedures, the *Simulated Minimally Competent Candidate* (SMCC) method for

the selected response format items and the *Examinee Paper Selection Method* for the essay portion, were combined to produce passing standard recommendations for each panelist. Because some sections of the TOEFL are adaptive and the test uses item response theory (IRT) to equate all examinee's performances to a common scale, it was desirable to use a standard setting procedure that was congruent with adaptive testing and IRT. The SMCC method is such a procedure. This method essentially asks each panelist to respond to a sample of items the way they imagine a minimally competent examinee would. Based upon those responses, a score is computed for the panelist that should represent the panelist's notion of minimal competence. If the items are already calibrated using IRT, the tests can even be given adaptively. Therefore, this method permits each rater to receive either different sets of items or identical sets of items. In this way, items can be administered to panelists in a manner similar to how an examinee would receive them on an actual test. The SMCC procedure produced a subtest score for the Listening Comprehension and Reading Comprehension subtests. It also provided a partial subtest score for Structure & Written Expression.

It seemed impractical to have the panelists attempt to write an essay in the way they imagine a minimally competent candidate would and then have the operational raters score it. The opportunity for the panelists to reconsider their ratings after discussion would be lost. Instead, the *Examinee Paper Selection* method (Hambleton, Jaeger, Plake, & Mills, 2000) was used which permitted the panelists to read the rubric descriptions, the elements of what constituted each point (0-6) on the rubric, and then read sample essay responses that corresponded to each point on the rubric. Panelists were asked to pick the response that, in their expert judgment, reflected the response of the single examinee with just enough English language skills to perform the job of an entry-level nurse safely and effectively. Panelists were permitted to use half points if they felt

that the minimally competent candidate would perform somewhere between two adjacent exemplars. This is consistent with the actual scoring process because two raters grade each essay and the average rating is used (Educational Testing Service, 2003). A conversion table was used to combine the selected rating with the partial subtest score for Structure & Written Expression that was generated using the SMCC procedure. The SMCC procedure and the *Examinee Paper Selection* method each contributed approximately 50% of the Structure/ Written Expression subtest.

Finally, there were three subtest scores for each panelist that could have ranged from 0 to 30. These subtest scores were combined into a total score (0 to 300) by summing the three section scaled scores and multiplying this sum by ten-thirds, effectively allowing each section scaled score to contribute equally to the total scaled score.

Adaptive Testing

Reading was administered as a fixed form test, but the Listening Comprehension and Structure sections were administered adaptively. That is, the difficulty level of an item presented to a candidate is dependent on the candidate's response to the immediate previous item and to the other previous items. A correct response to an item, for example, is followed by an item of greater difficulty; an incorrect response is followed by an item of lesser difficulty. In this way, a candidate receives a set of items maximally tailored to his or her overall ability in each of the two adaptive sections.

PowerPrep®

Each panelist was provided with a laptop computer that was preloaded with TOEFL PowerPrep software. PowerPrep contains two full-length, computer-adaptive editions of the TOEFL, drawing upon a pool of more than 1,200 items. The software does not provide a final score for the Structure & Written

Expression section, but instead it produces the lower bound of the Structure score, which essentially assumes that zero points were earned on the essay. Panelists were to combine this Structure score with their essay score through the use of a conversion table to produce a single score for the Structure & Written Expression section. For each section, the panelist's ability estimate was translated to a scaled score that could range from 0 to 30. Finally, a total score (0 to 300) for the panelist was obtained by summing the three section scaled scores and multiplying this sum by ten-thirds, effectively allowing each section scaled score to contribute equally to the total scaled score.

Selection of Raters

The composition (number, representativeness, and qualifications) of the standard-setting panel was a crucial element in establishing the validity and credibility of the standard. Twenty-five experts served on the standard-setting panel (Table 1). The panel, as a group, was intentionally made ethnically and linguistically diverse. Applicants were grouped by the following categories a) having previously taken the TOEFL exam, b) working with clients who speak languages other than English, c) supervising nurses who speak languages other than English, or d) working as nursing regulators. NCSBN further sorted applicants by selecting candidates from each of the most commonly spoken non-English languages in the U.S., and selecting representatives from all four NCSBN geographic regions. These experts, all female, were recruited by NCSBN to represent a range of professional perspectives and experiences. Collectively, 18 jurisdictions were represented on the panel: AK, CA, DC, FL, GA, HI, IA, IL, KS, LA, MA, MN, NJ, NC, OH, OR, TX, and VA.

Panelist Orientation and Training

The panelists were first provided with an overview of the goals and purpose of the study. It was explained

that a passing score was meant to reflect the level of English language proficiency necessary for entry-level nurses to perform important nursing tasks safely and effectively. It was clarified that the passing score was not the level of English language proficiency necessary to take the NCLEX examination—the focus of the study was on the job. Second, the panelists were led through an overview of the TOEFL computer-based test and the general process that was to be followed in arriving at the recommended passing score.

After the orientation, the panel was asked to identify the core tasks that all entry-level nurses needed to perform. It was important to agree on the scope of activity that was being considered before trying to assess how much English one needed to know to perform them. The list included: taking patient histories, conducting patient assessment, completing documentation, educating-training patients, taking orders, reporting, implementing safety practices, delegating, communicating, providing client service, and prioritizing responsibilities. This list was posted

Table 1. Panel Demographics (25 Panelists)		
	Number	Percent
Gender		
Female	25	100%
Panelist Selection Criteria		
Eligibility Criteria 1:		
Nurses who have taken the TOEFL	9	36%
Eligibility Criteria 2:		
Nurses who work with clients who speak a primary language other than English	7	28%
Eligibility Criteria 3:		
Clinical supervisors of nurses who speak a primary language other than English	5	20%
Eligibility Criteria 4:		
Nursing regulators, Nursing Educators, or Public Members	4	16%
Years of Nursing Experience		
1 to 5 Years	7	28%
6 to 10 Years	3	12%
11 to 15 Years	3	12%
16 to 20 Years	3	12%
More than 20 Years	5	20%
No Response	4	16%
Panelists Who Took TOEFL Test		
Yes	11	44%
No	14	56%
States Represented	18 (see text for the list of states)	

to serve as a frame-of-reference for the rest of the exercise.

Next, the panelists were instructed to imagine a nurse candidate who was educated outside the United States and in a language other than English. Furthermore, this imaginary candidate was seeking to become an entry-level nurse in the U.S. and just barely possessed the English proficiency necessary to be safe and effective as a nurse. Panelists were reminded that the focus was not on the examinee's nursing knowledge or skill, but rather on their English language skills.

The panelists were instructed to take the TOEFL examination on the computer at their table, however, they were to respond to each question as if they were the minimally proficient examinee that they had just imagined. For the multiple-choice questions, the panelists were instructed to select the answer that they believed the SMCC would choose. For the writing sample, the panelists were instructed to identify from a set of exemplar writing samples, the writing sample that reflects what the SMCC would be capable of producing.

RESULTS

Panel Recommendations

Before the quantitative results are summarized, it is important to note that the social dynamics among the participants and between the staff and participants was one of collaboration. It appeared that no panelist was reticent to provide their opinion and no individual or small subgroup dominated the discussions. It appeared that the results provided by the panel represent the panelists' true opinions regarding the minimum English proficiency required to practice nursing at the entry-level.

For each panelist, three subtest scores and a total test score were computed for their first round

judgments (Table 2) and their final second round judgments (Table 3). The panelists tended to indicate a higher standard was needed when they reconsidered their initial judgments. The mean score increased from 212 in the first round to 221 in the second round. While the mean and median values increased, the variability (standard deviation) of the panelists' judgments tended to decrease, indicating a greater degree of panelist consensus. This was expected, as previous research (Hurtz & Auerbach, 2003) indicates that group discussion of standard-setting judgment can result in reduced variability among panelist judgments and higher mean values.

In addition to considering the panel as a whole, the scores from participants who had previously taken the TOEFL (Table 4) and those who had not (Table 5) were considered separately. This was done because those panelists who had previously taken the TOEFL examination as part of the emigration or licensing process may have had a different perspective regarding minimum competence. Yet, the data did not support this hypothesis, as the mean scores of the group that had taken the TOEFL ($M=218$) and the mean scores of the group that had not ($M=223$) were not different to a statistically significant extent, $t(23)=0.721$, $p=0.48$, two tailed. Both of these groups tended to indicate after discussion that a higher standard was required than they thought in their initial judgment and both groups tended to show less variability in their post-discussion judgments. Furthermore, it was interesting to note that across subgroups, the mean scores for each of the three subtests (Listening, Writing-Structure, and Reading) were all very similar. Despite the general subtest agreement among the panelists, two panelists, #10 and #12, produced listening subtest scores that were noticeably lower than those of their colleagues. Both panelists raised their ratings after the discussion with their colleagues, but still remained the lowest subtest score.

Table 2. First Round Scores for All Panelists						
Panelist	Listening (0-30)	Structure Component (0-13)	Writing Component (0-6)	Combined Structure and Writing (0-30)	Reading (0-30)	Total (0-300)
P1	22	12	3.5	23	19	213.33
P2	22	10	4	23	26	236.67
P3	18	6	5	22	24	213.33
P4	24	9	5	25	23	240.00
P5	25	11	4	24	24	243.33
P6	24	6	3.5	18	23	216.67
P7	16	6	4	19	22	190.00
P8	15	2	4	15	19	163.33
P9	17	11	5	26	24	223.33
P10	13	7	3.5	19	22	180.00
P11	25	12	3	22	20	223.33
P12	10	6	4	19	24	176.67
P13	24	13	4.5	26	25	250.00
P14	23	9	4	22	19	213.33
P15	23	10	4	23	19	216.67
P16	20	10	3.5	22	24	220.00
P17	22	8	4	21	24	223.33
P18	21	3	4	16	18	183.33
P19	22	8	3.5	20	19	203.33
P20	22	9	4	22	21	216.67
P21	22	5	4	18	16	186.67
P22	26	13	3.5	24	21	236.67
P23	24	5	4.5	20	18	206.67
P24	20	11	3	21	19	200.00
P25	24	12	4	25	25	246.67
Mean (truncated)	20	8	3	21		212
Median (truncated)	22	9	4	22		216
Standard Deviation	3.94	3.02	0.53	2.87		22.77
Minimum	10	2	3	15		163.33
Maximum	26	13	5	26		250.00

- THE READABILITY OF NCLEX® EXAMINATIONS
 NCLEX® PSYCHOMETRIC RESEARCH BRIEF | VOL 1 | SEPTEMBER 2004

Table 3. Second Round Scores for All Panelists						
Panelist	Listening (0-30)	Structure Component (0-13)	Writing Component (0-6)	Combined Structure and Writing (0-30)	Reading (0-30)	Total (0-300)
P1	20	8	4	21	20	203.33
P2	22	10	4	23	23	226.67
P3	26	9	5	25	26	256.67
P4	24	9	4	22	22	226.67
P5	22	10	4.5	24	21	223.33
P6	24	10	3.5	22	23	230.00
P7	20	6	4	19	22	203.33
P8	24	8	4	21	20	216.67
P9	20	11	4.5	25	24	230.00
P10	18	8	3.5	20	22	200.00
P11	20	9	3.5	21	20	203.33
P12	15	7	4	20	24	196.67
P13	23	11	4	24	23	233.33
P14	23	9	4.5	24	22	230.00
P15	21	10	4.5	24	20	216.67
P16	20	10	4.5	24	24	226.67
P17	22	10	4.5	24	24	233.33
P18	25	9	4	22	25	240.00
P19	21	8	3.5	20	19	200.00
P20	22	8	4	21	21	213.33
P21	22	9	4	22	22	220.00
P22	26	11	3.5	23	25	246.67
P23	22	10	3.5	22	20	213.33
P24	20	8	3.5	20	20	200.00
P25	22	11	4	24	25	236.67
Mean (truncated)	21	9	4	22	22	221
Median (truncated)	22	9	4	22	22	223
Standard Deviation	2.40	1.29	0.41	1.73	1.95	15.74
Minimum	15	6	3.5	19	19	196.67
Maximum	26	11	5	25	26	256.67

Examination Committee Deliberation

The NCSBN Board of Directors charged the Examination Committee with developing a recommended minimum passing score for the TOEFL. The committee reviewed (1) the panel's recommendations in conjunction with (2) existing U.S. visa-screening requirements, (3) state licensing criteria, and (4) normative TOEFL performance data on people applying for a professional license (Table 6).

The individual recommended passing standards ranged from 197 to 257 with no drastic outliers. The difference between the mean and median was so small that it does not seem to reflect much difference in terms of language proficiency. Given that the group of panelists that had previously taken the TOEFL and those who had not, produced comparable scores, the committee felt strongly that the recommended standard should consider the opinions of the entire panel not just a subset of the panel. The mean score for the entire panel was 221 and that was the initial idea for the standard.

Additional discussion led the committee to consider the current U.S. visa-screening requirements for the different professions (Table 6). The current requirement for practical or vocational nurses is a TOEFL score of 197 and a score of 207 for registered nurses. However, NCSBN staff was unable to uncover any research or documentation to support those standards. Staff did find a standard setting study performed for Occupational Therapists and Physical Therapists. This study recommended a TOEFL score of 220 to be considered minimally competent. The Examination Committee considered the level of communication required for those jobs and concluded that entry-level nurses needed to have comparable communication abilities. This led the committee to revise their recommended standard to 220.

The Examination Committee then reviewed this standard in light of the standards used by some states as a licensing requirement. Many states use the U.S.

visa-screening requirements as their criteria, but there are some exceptions. The *2002 Profiles of Member Boards* identifies three states with different standards (Kansas 163, North Carolina 213, and Florida 217). The committee did not feel that the standards used by these states were better supported with research or rationale than their proposed standard. However it was reassuring that two of the three state standards were close to the committees current thinking, a standard of 220.

The committee also wanted to have a general idea regarding the impact their standard would have. The committee looked at the 2001-2002 test score information for the computer-based TOEFL (Table 6). TOEFL examinees who reported that they were taking the test to become licensed to practice their chosen profession are likely to be more similar to pool of internationally educated nurses in this study than would likely be the entire pool of TOEFL examinees. The subset of examinees seeking a professional license was further disaggregated by gender. This permitted the test score information to be better aligned with the demographic characteristics of the pool of internationally educated nurses, that is likely to contain more women than men.

To make a prediction regarding the impact of the proposed TOEFL standard of 220, a few scenarios were modeled. First it was assumed that the distribution of English language proficiency among internationally educated nurses taking the TOEFL was the same as the distribution of all TOEFL examinees applying for a professional license. Given that the mean and standard deviation for the population applying for a professional license was $\mu=229$, $s=42$, one would expect that 58% ($z=-0.214$) of these TOEFL examinees would pass. However, if the population were limited to female TOEFL examinees ($\mu=225$, $s=40$), one would expect that 55% ($z=-0.125$) of this group would pass. On the other hand, using the data reported by the Commission on Graduates of

- THE READABILITY OF NCLEX® EXAMINATIONS
 NCLEX® PSYCHOMETRIC RESEARCH BRIEF | VOL 1 | SEPTEMBER 2004

Table 4. First and Second Round Scores for All Panelists who have taken TOEFL (11 panelists)						
Panelist	Listening (0-30)	Structure Component (0-13)	Writing Component (0-6)	Combined Structure and Writing (0-30)	Reading (0-30)	Total (0-300)
ROUND 1 JUDGMENT						
P1	22	12	3.5	23	19	213.33
P4	24	9	5	25	23	240.00
P5	25	11	4	24	24	243.33
P7	16	6	4	19	22	190.00
P12	10	6	4	19	24	176.67
P14	23	9	4	22	19	213.33
P16	20	10	3.5	22	24	220.00
P20	22	9	4	22	21	216.67
P21	22	5	4	18	16	186.67
P22	26	13	3.5	24	21	236.67
P23	24	5	4.5	20	18	206.67
Mean (truncated)	21	8	4	21	21	213
Median (truncated)	22	9	4.	22	21	213
Standard Deviation	4.39	2.67	0.43	2.23	2.59	21.01
Minimum	10	5	3.5	18	16	176.67
Maximum	26	13	5	25	24	243.33
ROUND 2 JUDGMENT						
P1	20	8	4	21	20	203.33
P4	24	9	4	22	22	226.67
P5	22	10	4.5	24	21	223.33
P7	20	6	4	19	22	203.33
P12	15	7	4	20	24	196.67
P14	23	9	4.5	24	22	230.00
P16	20	10	4.5	24	24	226.67
P20	22	8	4	21	21	213.33
P21	22	9	4	22	22	220.00
P22	26	11	3.5	23	25	246.67
P23	22	10	3.5	22	20	213.33
Mean (truncated)	21	8	4	22	22	218
Median (truncated)	22	9	4	22	22	220
Standard Deviation	2.68	1.40	0.33	1.60	1.56	13.73
Minimum	15	6	3.5	19	20	196.67
Maximum	26	11	4.5	24	25	246.67

- THE READABILITY OF NCLEX® EXAMINATIONS
 NCLEX® PSYCHOMETRIC RESEARCH BRIEF | VOL 1 | SEPTEMBER 2004

Table 5. First and Second Round Scores for Panelists who have not taken TOEFL (14 panelists)						
Panelist	Listening (0-30)	Structure Component (0-13)	Writing Component (0-6)	Combined Structure and Writing (0-30)	Reading (0-30)	Total (0-300)
ROUND 1 JUDGMENT						
P2	22	10	4	23	26	236.67
P3	18	6	5	22	24	213.33
P6	24	6	3.5	18	23	216.67
P8	15	2	4	15	19	163.33
P9	17	11	5	26	24	223.33
P10	13	7	3.5	19	22	180.00
P11	25	12	3	22	20	223.33
P13	24	13	4.5	26	25	250.00
P15	23	10	4	23	19	216.67
P17	22	8	4	21	24	223.33
P18	21	3	4	16	18	183.33
P19	22	8	3.5	20	19	203.33
P24	20	11	3	21	19	200.00
P25	24	12	4	25	25	246.67
Mean (truncated)	20	8	3	21	21	212
Median (truncated)	22	9	4	21	22	216
Standard Deviation	3.53	3.27	0.59	3.28	2.71	24.07
Minimum	13	2	3	15	18	163.33
Maximum	25	13	5	26	26	250

- THE READABILITY OF NCLEX® EXAMINATIONS
 NCLEX® PSYCHOMETRIC RESEARCH BRIEF | VOL 1 | SEPTEMBER 2004

Table 5, continued

Panelist	Listening (0-30)	Structure Component (0-13)	Writing Component (0-6)	Combined Structure and Writing (0-30)	Reading (0-30)	Total (0-300)
ROUND 2 JUDGMENT						
P2	22	10	4	23	23	226.67
P3	26	9	5	25	26	256.67
P6	24	10	3.5	22	23	230.00
P8	24	8	4	21	20	216.67
P9	20	11	4.5	25	24	230.00
P10	18	8	3.5	20	22	200.00
P11	20	9	3.5	21	20	203.33
P13	23	11	4	24	23	233.33
P15	21	10	4.5	24	20	216.67
P17	22	10	4.5	24	24	233.33
P18	25	9	4	22	25	240.00
P19	21	8	3.5	20	19	200.00
P24	20	8	3.5	20	20	200.00
P25	22	11	4	24	25	236.67
Mean (truncated)	22	9	4	22	22	223
Median (truncated)	22	9	4	22	23	228
Standard Deviation	2.14	1.12	0.46	1.80	2.19	16.88
Minimum	18	8	3.5	20	19	200
Maximum	26	11	5	25	26	256.67

- THE READABILITY OF NCLEX® EXAMINATIONS
 NCLEX® PSYCHOMETRIC RESEARCH BRIEF | VOL 1 | SEPTEMBER 2004

Table 6. Information Considered by NCSBN's Examination Committee to Recommend a Minimum Standard on TOEFL	
Score¹	Source
Panel Recommendations	
218	Mean final recommendation of panelists who took TOEFL.
220	Median final recommendation of panelists who took TOEFL.
221	Mean final recommendation of all panelists.
223	Median final recommendation of all panelists.
223	Mean final recommendation of panelists who did not take TOEFL.
228	Median final recommendation of panelists who did not take TOEFL.
CGFNS' TOEFL Requirements for Visa Screening²	
197	Current Standard for Licensed Practical Nurses, Vocational Nurses, Clinical Laboratory Technicians, and Medical Technicians.
207	Current Standard for Registered Nurses, Speech Language Pathologists, Audiologists, Clinical Laboratory Scientists, Medical Technologists, and Physician Assistants.
220	Current Standard for Physical Therapists and Occupational Therapists.
State TOEFL Standards³	
163	Kansas
213	North Carolina
217	Florida
Population Means and Standard Deviations for Computer-based TOEFL 2001-2002 Examinees⁴	
214 (47)	Mean score of all examinees taking the computer-based TOEFL (N=572,394)
225 (40)	Mean score of female 2001-2002 TOEFL examinees that applied for any type of professional license. (N=21,187)
229 (42)	Mean score of all 2001-2002 TOEFL examinees that applied for any type of professional license. (N=34,721)
235 (44)	Mean score of male 2001-2002 TOEFL examinees that applied for any type of professional license. (N=13,283)
CGFNS Validity Study Sample Mean and Standard Deviation⁵	
237 (19)	The CGFNS TOEFL sample was based on the written examination, not the CAT examination. The written examination scores were converted to CAT scores via the following formula $CAT = (Written - 273.9) * 0.769$. This formula was based on a conversion table found on page 13 of the TOEFL 2003-04 Information Bulletin for Computer-based and Paper-based Testing.

¹ TOEFL scores can range from 0 to 300.

² This information comes from the CGFNS website as of Jan 22, 2004.

³ This information comes from the Profile of Member Boards 2002 (NCSBN, 2003).

⁴ Numbers of examinees are based on those who responded to a question about their group membership.

⁵ Based on the Commission on Graduates of Foreign Nursing Schools' Validity Study April 1999 through March 2000.

Foreign Nursing Schools (CGFNS, 2000) in their validity study, one could get a better idea of the typical distribution of English language proficiency for internationally educated nurses taking the TOEFL. Using only the people reported in that study who were in the 1999 or 2000 TOEFL cohort, an estimate for the population of nurses was derived ($\mu=237.5$, $s=19$)². Using this population, one would expect 82% ($z = -0.921$) of them to pass.

The Examination Committee considered the impact predictions and agreed that a standard of 220 on the TOEFL was appropriate to demonstrate the minimum degree of English proficiency necessary to be a safe and effective, entry-level nurse. Correspondingly, a score on of 560 on the paper version of the TOEFL would be considered equivalent.

DISCUSSION

The purpose of this study was to arrive at a recommended passing score on the TOEFL that represented the level of level of English language proficiency believed necessary to perform important entry-level nursing tasks safely and effectively. The Examination Committee was asked to make a policy decision after being informed with the appropriate types of information. The committee did this after giving consideration to a broad spectrum of information. As a result, there is now a recommended passing standard for entry-level nursing that can be supported by carefully documented and well-designed procedures.

Limitations

Typically, there are some shortcomings that are inherent in tests that are related to licensure and certification testing. Test developers are often restricted in the types of data that they can collect to verify the standard. In practice, boards only license or certify people that are believed to be competent. Were they believed to be incompetent, it would be unethical to license or certify them. Because these people come only from the upper end of the ability continuum, there are sampling problems related to attempting to establish the predictive validity of the standard. Therefore, this type of predictive validity is not normally demonstrated for certification and licensure tests.

Future Activities

Now that the standard has been set, the question is now how many of the boards of nursing will use this standard as a legal requirement for licensure? Also, the adoption of these standards for visa screening purposes is also of interest. Because the adoption and implementation of this standard rests with governmental entities, NCSBN's role is one of providing information and documentation about the standard. In the future, NCSBN intends to provide recommended standards for other English Proficiency examinations as well. This will provide boards of nursing and candidates with more choices in tests and test providers.

²The CGFNS TOEFL sample was based on the written examination, not the CAT examination. The written examination scores were converted to CAT scores via the following formula $CAT = (Written - 273.9) * 0.769$. This formula was based on a conversion table found on page 13 of the TOEFL 2003-04 Information Bulletin for Computer-based and Paper-based Testing.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Cizek, G. J. (2001). Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Commission on Graduates of Foreign Nursing Schools. (2000). Commission on Graduates of Foreign Nursing Schools Validity Study April 1999 through March 2000. Unpublished statistical report.

Educational Testing Service. (2003). TOEFL 2003-04 Information bulletin for computer-based and paper-based testing. Princeton, NJ: Educational Testing Service.

Hambleton, R.K., Jaeger, R.M., Plake, B.S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24, 355-366.

Hurtz, G.M., & Auerbach, M.A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, 63, 584-601.

National Council of State Boards of Nursing, Inc. (2003). Profile of Member Boards 2003. Chicago, IL: National Council of State Boards of Nursing, Inc.

A special thanks to the members of the NCSBN Examination Committee who debated and deliberated to produce a well conceived standard. Also, thanks to Anne Wendt, RN, PhD, CAE, who recommended many of the specifications for the selection of panelist.



National Council of State Boards of Nursing
111 E. Wacker Drive, Suite 2900
Chicago, IL 60601-4277 USA
312.525.3600 | International Calls: +1.312.525.3600
Testing Services: 312.525.3750 | www.ncsbn.org